

GENERATIVE AI AND CRIMINAL LIABILITY: DEEPPFAKE-ENABLED FRAUD AND ATTRIBUTION IN 2026

Sergio Pommier Gallo, PhD.

ORCID iD: 0009-0000-7387-3943

Affiliation: Stanford University

Website: <https://sergiopommiergallophd.blogspot.com>

Abstract

This study examines the emergence of generative artificial intelligence as a structural enabler of deepfake-based fraud and extortion within contemporary criminal systems. Moving beyond traditional evidentiary paradigms, the research analyzes how synthetic media disrupts attribution, evidentiary reliability, and doctrinal interpretations of mens rea. Using an interdisciplinary framework integrating criminal law, digital forensics, and criminological analysis, the study identifies patterns of deception, operational scalability, and attribution ambiguity. The findings support the hypothesis that generative AI functions as an evidentiary distortion mechanism that complicates, but does not eliminate, individual criminal responsibility. The article proposes a doctrinal model of algorithmic attribution that distinguishes between synthetic output and human agency, reinforcing evidentiary safeguards necessary for due process in technologically mediated criminal environments.

Keywords: artificial intelligence; deepfakes; criminal liability; digital evidence; attribution; fraud; cybercrime.

1. Introduction

The rapid development of generative artificial intelligence has introduced a new paradigm in the production of synthetic audiovisual content. By 2026, deepfake technologies have evolved from experimental tools into operational instruments used in fraud, extortion, reputational attacks, and financial deception. These technologies enable the fabrication of highly convincing representations of

individuals, including voice cloning and real-time video manipulation, thereby significantly expanding the operational capabilities of criminal actors.

Traditional criminal law frameworks are grounded in assumptions of evidentiary stability and identifiable authorship. However, deepfake technologies undermine these assumptions by decoupling observable evidence from verifiable human conduct. This creates a critical tension between technological capability and doctrinal requirements for attribution and intent.

Research Question: How does generative AI-based synthetic media affect attribution, evidentiary reliability, and criminal liability in contemporary legal systems?

Hypothesis (H1): Deepfake-enabled criminal conduct introduces structural distortions in evidentiary evaluation, requiring enhanced doctrinal safeguards without altering the foundational principles of individual culpability.

2. Theoretical Framework

This study integrates criminal law doctrine with theories of technological mediation, digital deception, and evidentiary reliability.

The concept of synthetic evidence is defined as digitally generated content that simulates reality without direct correspondence to actual events. Unlike manipulated evidence, synthetic evidence may be entirely fabricated without any original referent.

Algorithmic attribution is introduced as a doctrinal concept referring to the process of linking human intent and agency to outcomes generated or mediated by artificial systems.

The framework distinguishes three layers:

- (1) generation (algorithmic creation),
- (2) deployment (human use and dissemination),
- (3) intent (criminal purpose).

This layered model preserves the centrality of human agency while recognizing the mediating role of technology.

3. Methodology

The study employs a qualitative doctrinal and criminological methodology.

Data sources include: legal scholarship; cybercrime reports; forensic analysis literature; documented case patterns; and interdisciplinary research on AI and law.

Analytical approach: doctrinal analysis combined with criminological pattern recognition and forensic reasoning.

Variables analyzed: attribution complexity; evidentiary reliability; intent inference; technological mediation; procedural safeguards.

Evidence classification:

High = peer-reviewed legal and scientific literature

Medium = institutional reports and documented case patterns

Inferential = doctrinal synthesis and criminological inference

4. Results

The analysis identifies four principal dynamics:

(1) Attribution Dilution

The presence of synthetic media complicates the identification of the actor responsible for content creation and dissemination.

(2) Evidentiary Instability

Digital artifacts can no longer be presumed authentic without technical validation.

(3) Scalability of Deception

Generative AI enables large-scale production of fraudulent content with minimal cost.

(4) Psychological Credibility

Deepfakes exploit cognitive biases, increasing victim susceptibility.

Operational Typology:

Voice Cloning Fraud — financial deception — low detection threshold

Synthetic Video Extortion — coercion — high psychological impact

Identity Simulation — reputational harm — attribution complexity

Hybrid Attacks — multi-platform deployment — rapid dissemination

These patterns demonstrate that deepfake-enabled crime is structurally distinct from traditional fraud.

5. Discussion

The findings support the hypothesis that generative AI introduces evidentiary distortion rather than doctrinal transformation.

Mens rea remains the central criterion for criminal liability. However, the inference of intent must rely on convergent evidence beyond synthetic artifacts. Communications, financial transactions, access logs, and behavioral patterns become essential corroborative elements.

The study rejects any deterministic interpretation of algorithmic outputs. Artificial intelligence does not generate intent; it mediates representation. Therefore, liability must remain anchored in human conduct.

The concept of algorithmic attribution provides a structured approach to resolving evidentiary ambiguity. By separating technological generation from human intent, courts can preserve doctrinal coherence while adapting to technological realities.

6. Limitations

The study is limited by reliance on secondary sources and the relative novelty of large-scale judicial treatment of deepfake evidence.

Additionally, rapid technological evolution may alter the characteristics of synthetic media, requiring continuous doctrinal adaptation.

Future research should include empirical case analysis, forensic benchmarking, and quantitative assessment of detection accuracy.

7. Conclusions

Generative AI represents a significant transformation in the evidentiary environment of criminal law.

Deepfake technologies introduce structural challenges in attribution, reliability, and procedural fairness. However, they do not alter the normative foundations of criminal liability.

The proposed model of algorithmic attribution offers a robust framework for distinguishing between synthetic outputs and human agency.

Effective legal responses require:

- (1) enhanced evidentiary standards
- (2) forensic validation protocols
- (3) procedural transparency
- (4) interdisciplinary integration between law and technology

Absent these safeguards, the risk of wrongful attribution and evidentiary distortion will increase substantially.

References (APA 7th)

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war. *Foreign Affairs*.

Greene, J., & Cohen, J. (2004). For the law, neuroscience changes nothing and everything. *Philosophical Transactions of the Royal Society B*, 359, 1775–1785.

<https://doi.org/10.1098/rstb.2004.1546>

Kerr, O. S. (2005). Digital evidence and the new criminal procedure. *Columbia Law Review*.

Pardo, M. S., & Patterson, D. (2013). *Minds, brains, and law*. Oxford University Press.

United Nations Office on Drugs and Crime (UNODC). (2023). Cybercrime and digital fraud report.

References (Chicago Style)

Chesney, Robert, and Danielle Citron. 2019. "Deepfakes and the New Disinformation War." *Foreign Affairs*.

Greene, Joshua, and Jonathan Cohen. 2004. "For the Law, Neuroscience Changes Nothing and Everything." *Philosophical Transactions B* 359: 1775–1785.

Kerr, Orin S. 2005. "Digital Evidence and Criminal Procedure." *Columbia Law Review*.

Pardo, Michael S., and Dennis Patterson. 2013. *Minds, Brains, and Law*. Oxford: Oxford University Press.

United Nations Office on Drugs and Crime. 2023. *Cybercrime and Digital Fraud Report*.

License

Creative Commons Attribution 4.0 (CC BY 4.0)

Version

Preprint (submitted manuscript, not peer-reviewed)